DOCUMENT RESUME

ED 126 120 95 TM 005 375

AUTHOR Roper, Susan Stavert; And Others
TITLE A Pilot Test of Collegial Evaluation for Teachers. Research and Development Memorandum No. 142.
INSTITUTION Stanford Univ., Calif. Stanford Center for Research and Development in Teaching.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO SCRDT-RDM-142
PUB DATE May 76
CONTRACT NE-C-00-3-0062
NOTE 32p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Classroom Observation Techniques; Evaluation Criteria; Feedback; *Helping Relationship; Pilot Projects; Self Evaluation; Student Evaluation of Teacher Performance; Student Teachers; *Teacher Evaluation; *Teacher Improvement; *Teacher Participation; Teachers; *Teamwork; Test Construction
IDENTIFIERS *Collegial Evaluation

ABSTRACT

When public pressure mounts for teacher accountability, current methods of evaluating teachers are widely regarded as inadequate. Teachers often feel that evaluation is hasty, arbitrary, and threatening; more important, it gives them little practical help in improving their performance. This paper describes a pilot test of a new collegial evaluation program that emphasizes the improvement of classroom teaching. Working in pairs, teachers select their own criteria, observe each other in the classroom, give each other feedback, and develop plans for improvement. The program also provides for self-assessment and assessment by students to be incorporated into the overall evaluation. Thirty teachers and teacher trainees drawn from a variety of teaching situations participated in the pilot test. On the whole, the results were promising: teachers reacted favorable to collegial evaluation; they were able to adapt the program to their own needs when necessary; and they gained new ideas for improvement from it. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED126120

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM | SP
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

SCRDT

Stanford Center for Research and Development in Teaching

SCHOOL OF EDUCATION STANFORD UNIVERSITY

Stanford Center for Research and Development in Teaching
School of Education, Stanford University
Stanford, California

Research and Development Memorandum No. 142

A PILOT TEST OF COLLEGIAL EVALUATION FOR TEACHERS

Susan Stavert Roper, Terrence E. Deal,
and Sanford M. Dornbusch

May 1976

Published by the Stanford Center for Research and Development in Teaching, supported in part as a research and development center by funds from the National Institute of Education, U. S. Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. NE-C-00-3-0062.)

Introductory Statement

The mission of the Stanford Center for Research and Development in Teaching is to improve teaching in American schools. Current major operations include three research and development programs--Teaching Effectiveness, The Environment for Teaching, and Teaching and Linguistic Pluralism--and two programs combining research and technical assistance, the Stanford Urban/Rural Leadership Training Institute and the Hoover/Stanford Teacher Corps Project. The ERIC Clearinghouse on Information Resources is also a part of the Center. A program of exploratory and related studies provides for smaller studies not part of the major programs.

This report describes the experiences of teachers and interns who participated in a program to improve teaching through collegial evaluation. Both the evaluation program and pilot test reported here were designed and carried out by the Environment for Teaching Program. A manual that provides step-by-step directions for implementing the collegial evaluation program is in preparation.

Contents

Abstract iv

The Collegial Evaluation Program 4

The Pilot Test 5

Choosing a Partner 6

Selecting Evaluation Criteria 8

Observations 11

Conferences 12

Self-Assessment and Student Questionnaire 17

Self-Assessment on Selected Criteria 20

The Improvement Plan 21

Conclusions 24

References 26

Abstract

While public pressure mounts for teacher accountability, current methods of evaluating teachers are widely regarded as inadequate. Teachers often feel that evaluation is hasty, arbitrary, and threatening; more important, it gives them little practical help in improving their performance.

This paper describes a pilot test of a new collegial evaluation program that emphasizes the improvement of classroom teaching. Working in pairs, teachers select their own criteria, observe each other in the classroom, give each other feedback, and develop plans for improvement. The program also provides for self-assessment and assessment by students to be incorporated into the overall evaluation.

Thirty teachers and teacher trainees drawn from a variety of teaching situations participated in the pilot test. On the whole, the results were promising: teachers reacted favorably to collegial evaluation; they were able to adapt the program to their own needs when necessary; and they gained new ideas for improvement from it.

A PILOT TEST OF COLLEGIAL EVALUATION FOR TEACHERS

Susan Stavert Roper, Terrence E. Deal, and Sanford M. Dornbusch

Schoolteachers and administrators alike are experiencing pressures to improve classroom teaching as the result of a general movement toward greater "accountability" in our educational system. In formulating policies to institute accountability, state legislators have assumed that required evaluation procedures will automatically result in better teaching. Unfortunately, nearly everyone in the educational field agrees that the evaluation of teachers is poorly done and gives teachers little practical help in improving their performance. But here the agreement ends. Some feel that improving principals' skills in evaluating teaching performances is the answer. Others, feeling that the principal is overworked, would bring in outside evaluators to inspect classroom teaching. Still others would shift the focus from teaching performances to educational outcomes measured through achievement tests or behavioral objectives.

In the last decade we have studied evaluation processes in many different organizations, including not only schools but an assembly line, a physics research team, hospitals, a Roman Catholic archdiocese, university faculties, a student newspaper, and even a football team. From these studies we have developed a general model of evaluation that consists of six steps: (a) assigning goals, (b) setting criteria or standards, (c) making observations (sampling performance), (d) appraising performance, (e) communicating appraisals (providing feedback), (f) planning a program for improvement. These steps are interdependent; a weakness in any one lessens the contribution that evaluation can make to improving job performance (Dornbusch and Scott, 1975).

In our educational research we have gathered information on evaluation processes from 600 teachers and 33 administrators. This information was sufficient to convince us that weaknesses in one, two, or even all six evaluation steps were common in schools. For example, in one

---

A shorter version of this paper will appear as "Collegial Evaluation: Does It Work?" in Educational Research Quarterly, Spring 1976.

study that was part of this research, about half of the teachers reported that they did not know what criteria were used to evaluate them, or that the evaluation criteria were too vague to be meaningful (Thompson, Dornbusch, and Scott, 1975). They complained that observations of their classroom teaching often amounted to no more than infrequent quick peeks into the classroom by the principal. It is not surprising, then, that teachers are very anxious about being evaluated and do not believe that evaluation helps them improve their teaching.

But principals are not necessarily to blame for these shortcomings. Often they are too busy with administrative duties to spend adequate time observing teachers or providing them with useful feedback. Also, because of their formal supervisory position, their evaluations often seem threatening to teachers. And certainly the sense of threat has been amplified by the punitive implications of accountability legislation. As a consequence, many administrators feel more comfortable using students' test scores as an indirect means of assessing teacher performance.

Some recent evaluation programs stress the importance of student learning as an indicator of successful classroom teaching. But the emphasis on student outcomes creates problems of its own. For one thing, student variability is often so great that student outcomes may tell us less about teachers than about students, or even evaluators. More important, teachers have justifiably asked: How will these results help us improve our teaching? As the sports cliché goes, "Knowing the score doesn't help the team improve for the next game." In sum, evaluation programs that rely on the principal as the sole evaluator or on student outcomes as a means of assessing teacher quality will do little to improve teaching.

Our investigation has convinced us that at a minimum, an evaluation program aimed at improving classroom teaching must have three characteristics: (1) it must not have punitive implications; (2) it must designate evaluators to supplement the principal's evaluation; and (3) it must focus on teaching performance. In this paper we propose collegial evaluation as a strategy for satisfying these three criteria.

Collegial evaluation is advocated in many professional organizations as a means of both maintaining standards and improving performance. In fact, professionals derive much of their status from the fact that they have formally assumed responsibility for evaluating one another through their occupational associations. Clearly, a program in which teachers evaluate one another poses some difficulties. Teachers have had almost no experience in formally evaluating teaching, and often they consider their classroom a private domain, out of bounds to others except the principal. Despite these difficulties, our studies of open-space classrooms and team teaching revealed that reducing the isolation of classroom teachers had unexpected positive results. Since teaching performance was more visible under these conditions than in the traditional classroom, teachers viewed evaluation of their teaching by colleagues as more legitimate. As these evaluations were exchanged, teachers developed more respect for the ability of their colleagues to make sound judgements. As a result, they were more willing to have colleagues evaluate their teaching (Marram, Dornbusch, and Scott, 1972). From these findings, we reasoned that if teachers were given the opportunity to observe one another, they could give each other useful feedback. They would thus become more willing to evaluate each other in the future and would be able to use these regular evaluations to improve their classroom teaching. Of course, not all teachers will be happy with collegial evaluation. But since teachers are so disenchanted with the present hierarchical structure of evaluation, they may be receptive to a new approach in which they conduct their own evaluations.

We have developed a program of collegial evaluation for teachers that emphasizes evaluation as a means to improve classroom teaching. Since the spring of 1974 over 150 teachers have been introduced to the program through workshops. Thirty teachers participated in the pilot test. The teachers worked in pairs, selecting criteria, observing each other, providing feedback based on their observations, and helping one another develop specific plans to improve their teaching. In addition, teachers administered a student questionnaire and completed two types of self-assessment to gain more information to integrate into their improvement plans.

The results of this test suggest that our collegial evaluation program can help both experienced teachers and teacher trainees improve their teaching. This memorandum describes the experiences teachers had in each step of the collegial evaluation program in the hope that other teachers and administrators will be encouraged to give collegial evaluation a try.

The Collegial Evaluation Program

As we have mentioned, in our collegial evaluation program teachers work in evaluation partnerships to improve the quality of their teaching. The program also provides for self-assessment and student assessment. The entire sequence of self- and student assessment, observations, and conferences requires ten to twelve hours spread over a month or two. The program is flexible and can be implemented by an entire faculty, a department, a teaching team, or any two interested teachers. We are presently preparing a manual for teachers containing all the directions and forms needed to implement the program. In addition, the manual will explain the rationale for each step, incorporate examples of successful practices from other teachers, and offer suggestions to help teachers get the greatest benefit from this experience. The collegial evaluation program consists of seven interrelated steps:

1. Choosing a partner. This partnership between two teaching colleagues is the heart of the program of professional development.
2. Selecting evaluation criteria. In some schools teaching standards have been defined specifically enough to serve as a guide for evaluation. We help by providing examples of criteria used by other teachers.
3. Self-assessment. Each teacher completes a self-evaluation form, which is based partially on the criteria selected as well as on a questionnaire given to students.
4. Student assessment. A questionnaire is provided to get important feedback from students.
5. Observations. Observing a colleague's teaching and being observed in turn is the crucial step. We have developed forms for making

observations based on the selected criteria and tips to improve observational skills.

6. Conference on observations. After each set of observations, the collegial pair holds a conference. The purpose of the conference is to report observations and develop plans for improvement in appropriate areas. The program specifies a structure for the conference so that each teacher knows how to proceed.
7. The improvement plan. A final conference is held to pull together observations by colleagues, self-assessments, and student assessments. Once again, the structure of the conference is specified and some suggestions are provided for developing a long-term program for improvement.

The Pilot Test

Our pilot test was designed to serve several purposes: (1) finding out how teachers would react to collegial evaluation; (2) helping to identify the unforeseen problems that inevitably arise in any new venture; (3) discovering whether or not teachers would adapt the program to fit their own needs, and if so, in what ways; and (4) helping us improve the program. Of course, the critical question was, Will the collegial evaluation program work?

Although the pilot test sample included only 30 teachers, they were deliberately drawn from a variety of teaching situations. The situations included (1) different subject areas, (2) different grade levels (K-12), (3) suburban and inner-city schools, (4) open-space and self-contained classrooms, and (5) teachers with varying levels of experience, including both teacher interns and credentialed teachers.

We worked with two groups of teachers: teachers in an elementary school serving a California suburban community, and teacher trainees in the Stanford University teacher intern program. The interns were assigned to junior high schools and high schools on the San Francisco Peninsula and in San Jose, California. They were teaching numerous subjects including natural sciences, social studies, English, art, music, physical education,

and languages. Some were working in upper-middle-class schools, and others in predominately Black or Chicano inner-city schools. The elementary teachers worked in a school that has been architecturally designed to permit alternative instructional approaches in its various open-space "pods". The evaluation partners in this school were members of the same teaching team. The secondary teachers, all interns, had only a few months of teaching experience, while the elementary teachers varied in experience from two years to over fifteen. In sum, the diversity of teachers, students, and settings in the pilot test allowed us to determine whether the collegial evaluation program would work across a variety of different situations.

These teachers participated in all stages of the collegial evaluation program. Their experiences with various aspects of the program are summarized below, along with their comments and suggestions.

Choosing a Partner

As it happened, partnerships were formed quite differently among the elementary school teachers and the teacher trainees. The teacher trainees were free to choose their own partners, and they usually did so by common agreement on the basis of friendship or proximity. In the elementary school the principal assigned partners, and all of the partners had worked together previously in open-space classroom teams. The teachers were generally satisfied with the principal's assignment, but most teachers as well as interns felt that collegial evaluation participants should be able to select their own partners. All were skeptical of random selection as well.

There was some disagreement, however, about the criteria that should be used in selecting partners. Some stressed previous friendship. As one teacher said, "Working with a fellow teacher on this program required a lot of respect and trust. You've got to really like one another. It's almost like a marriage--only if you like someone can you be honest." Others emphasized the importance of choosing a partner from the same subject area and/or grade level to maximize the relevance of feedback. One intern, however, argued that she learned a great deal from observing a colleague in another subject area. Her field was English; her colleague

was in biology. She said that she was immediately impressed by the number and variety of materials available to students in the biology lab and realized for the first time how meager her English classroom materials were.

Some interns felt that teaching experience should be a critical factor in the selection of partners. They felt that a more experienced teacher would not take their criticisms or suggestions seriously, and that they would be hesitant to express their own fears of inadequacy to an "old pro." Other interns reported, to the contrary, that their partnership was limited by an insufficient experience base from which to generate "well-seasoned" suggestions for improving teaching. By contrast, the elementary teachers did not even mention teaching experience as a factor in the selection of partners. Levels of experience did not affect the quality of feedback or mutual respect within the elementary school group.

Both the interns and the elementary teachers agreed that partners should share a similar educational philosophy. The interns were particularly adamant on this point, maintaining that they tended to ignore criticism from someone whose views were radically different from their own. An illustration of the importance of educational philosophy came from a pair of interns who realized after observing each other that they both needed to become more directive and firm with their students. Although their supervisors had previously mentioned that they were losing control of the class, the interns had attributed this criticism to a philosophical conflict between them and their supervisors. They therefore made no attempt to change their teaching behavior. However, when they received similar feedback from someone they regarded as more sympathetic to their views, they were willing to take the criticism more seriously. Both wanted an "open," "trusting" classroom environment, but what they saw in their respective observations was "chaos." As one of the partners said:

> We have a philosophical stance that makes each of us uncomfortable with the role of "authority" figure in our classrooms. We are both searching for ways to make learning happen without crushing spirits, damaging self-concepts, or belittling individuals. It is our shared ideals that make it easy for me to accept Bill's criticism--I cannot reject it as being in disagreement with my fundamental beliefs.

Above all, the critical element in successful collegial evaluation is mutual respect. All participants agreed that respect for their partner's ability was more important than friendship, teaching experience, philosophy, or subject area as a basis for a successful collegial relationship.

Selecting Evaluation Criteria

The process of selecting evaluation criteria consists of five steps: (1) the two teachers identify the pool of possible criteria using such sources as school goals, accountability guidelines, recent research, and their own philosophy; (2) each teacher makes a list of four or five criteria and exchanges lists with his or her partner; (3) the two teachers agree on a list of four or five criteria; (4) the two teachers review the list to make sure each criterion is specific and observable; and (5) the criteria are listed on the observation form.

According to both the interns and the elementary teachers, selecting criteria was clearly the most difficult step in the collegial evaluation process. The main problem was developing criteria that were specific enough to be observable but still significant enough to reflect important aspects of teaching performance. For example, the criterion "ability to write clearly on the blackboard" would have been specific, but teachers were more interested in focusing on broader areas such as rapport with students and ability to motivate students. Criteria for observing rapport and ability to motivate were much more difficult for them to develop.

Teachers also reported difficulty in selecting criteria that could be applied to the actual situations in which observation took place. Some teachers had difficulty because their criteria were appropriate for a different instructional activity than the one they observed. Others selected criteria that focused on too many activities simultaneously.

Partners usually were able to agree on a list of criteria. Although the program provides an option whereby the two individuals may use completely different criteria, no one took this option in the pilot test.

As the partners developed their list, the major source of disagreement was the effort to define the criteria in specific, observable terms. As noted, part of the problem was caused by vague and ambiguous criteria that did not lend themselves to clear definition or observation. For example, teachers who selected "uses communication skills effectively," "degree of engagement by students," or "response of class to lesson" spent much of their meeting trying to decide what they meant by "communication skills," "engagement," or "response."

Although trying to select important yet observable criteria was difficult, teachers did not find the task boring or unproductive. As one said, "Selecting useful criteria forced me to clarify my own educational philosophy. I had to decide what was really important to me and then try to operationalize my goals so they could be observed."

Some of the better criteria were specific to certain subject matter. For example, two physical education teachers agreed that ensuring the physical safety of students was an important criterion for successfully teaching a tumbling lesson. During the lesson the observing teacher noted that although the tumbling mats had been carefully arranged, several students had not tied back their hair and were chewing gum during the practice session--both violations of safety rules. Similarly, two music teachers were able to give each other excellent feedback using criteria such as "time limit per piece," "explanation of the warm-up period," "presentation of rehearsal objectives," and "discussion of stops made during rehearsal."

Also useful were criteria that focused attention on specific mannerisms or behavior patterns of teachers. Watching for "any distracting speech mannerisms or gestures," a teacher discovered that her partner ended almost every sentence with a tentative "OK?" Until the first conference the teacher was totally unaware that she had this disturbing habit. One pair who taught in an open-area classroom listed "teacher mobility around the pod" as a criterion. Their observations revealed that one teacher was constantly moving around the classroom while the other was not moving enough to supervise students adequately.

Many teachers had difficulty making the leap from identifying a general area for observation and potential improvement to developing criteria that could be used to assess teaching performance in that area. But others were quite successful in developing appropriate criteria. For example, some teachers selected "motivates students to participate in discussion" as a general area and came up with "number of times teacher responded with a positive statement," "number of negative comments by teacher," "average length of time teacher waited for an answer," "number of students who were called on to answer a question," and "terms teacher used to praise or sanction students" for specific criteria. These teachers were able to learn about some of their specific behaviors that enhanced or hampered student motivation. For example, one teacher learned that the words she used to praise students were too dramatic when her partner observed that students dismissed her praise as unrealistic. She vowed to delete "fantastic" and "terrific" from her vocabulary except for "truly fantastic" responses. Another teacher learned that she habitually called only on students seated in the front rows. She decided to rotate seating positions weekly to enhance the opportunity of all students to participate. Overall, teachers recommended that in developing criteria the two partners should discuss how they would measure and observe a criterion before agreeing to use it.

There was some debate among teachers about the extent to which the criteria should reflect areas of potential weakness rather than areas of strength. Most agreed that potential weaknesses should be the basis for at least some of the criteria, since the purpose of collegial evaluation is improvement, not just reinforcement. The process of identifying areas of teaching weakness was more difficult for the elementary school teachers than for the interns. One teacher suggested that early observation of a teacher reputed to be exemplary might help in identifying one's own problem areas. It would be helpful, this teacher said, to have an opportunity to compare your own teaching performance with that of another teacher prior to deciding on criteria.

An extremely productive technique for generating criteria was to distribute the student questionnaire before selecting criteria. Our collegial evaluation manual will be revised to incorporate this finding. Some

teachers distributed the student questionnaire before meeting with their partner to select criteria. When one pair found that many students did not understand a teacher's directions, "clarity of directions" became the first criterion on their list.

In summary, teachers in the pilot test suggested that the evaluation partners decide together how they would observe a criterion before including it on their list. In selecting criteria they learned that vague, ambiguous, and global terms were not useful guidelines for observation. Criteria related to a particular subject or grade level were often helpful. Teachers found that the responses from the student questionnaire helped them generate criteria. Although they agreed that selecting criteria was the most difficult step in the collegial evaluation program, they thought it was worth the effort. Selecting criteria helped them decide not only where they needed to improve but what was most important to them as educators.

Observations

All teachers in the pilot test observed one another for two classroom periods, the observation time the program requires. Some participants felt that two periods did not provide enough time for observation. However, often these teachers had selected too many criteria or had selected criteria that were not applicable to the classroom situation they observed. More observations are certainly desirable, but since the program is designed to minimize inconvenience to teachers and administrators, it is preferable to take steps to make two observation periods sufficient. Selecting criteria appropriate for the classroom session is one example.

A number of interns selected different classes and/or subject areas for their second observation in order to learn whether their strengths and weaknesses were the same across different subjects and classrooms.

Teachers in the pilot tests reported that they learned as much from observing as from being observed. Many related a host of new teaching techniques acquired in their role as observer. In one open-space classroom the teachers switched students for observation. One teacher observed the other teaching a lesson to her students and vice versa.

Both teachers reportedly learned from this trade-off. One remarked that she had not realized a certain group of students never participated in her class until she sat at the back and watched them being taught by someone else.

The interns also reported benefits from observing. Under normal circumstances, interns rarely have the opportunity to see anyone teach a class other than their master teacher. As one intern put it, "By seeing other interns you get to see yourself with regard to your peer group--it is reassuring to know that you are not the only one making mistakes."

All participants liked the exposure to other methods of instruction and teaching styles. Teachers rarely have a chance to observe one another teaching--particularly if they are in self-contained classrooms. But even the teachers in the open-space school said that under usual conditions, they were too busy to observe their teammate adequately. Collegial evaluation gave them the chance not only to observe but to focus their observation using specific criteria.

The quality of feedback exchanged in the conferences was largely dependent on the quality of observations. The best observers were those guided by a few specific criteria that were appropriate to the particular activity they observed. They learned more from their observations and were better able to offer their partner concrete and useful information.

Conferences

Conferences require the ability to give constructive criticism without damaging egos or destroying long-term relationships. As our collegial evaluation program specifies, teachers in the pilot test exchanged feedback on three occasions: after each of the observation periods and at the wrap-up conference. In addition, they rated their strengths or weaknesses for each of the shared criteria on the self-evaluation form, which is similar to the observation form, making it easy to compare the two evaluations. In every case, participants were harder on themselves than their colleagues were.

The interns were much more willing than the elementary teachers to give low ratings to their colleagues and to give critical feedback on the observation form. Interns, by definition, are "people learning the skills of teaching," while certificated teachers (theoretically at least) already possess these skills. From this perspective, it is not surprising that interns were more comfortable offering written criticism than the elementary school teachers. During the conferences, however, teachers exchanged criticism and did more than pat one another on the back. Although they were reluctant to write down their negative comments, they were usually quite candid in their conferences.

An important purpose of the conferences is to develop specific strategies for improvement. Since the elementary school teachers worked together in the same classroom area, many of them identified problems that could be worked on cooperatively. For example, one pair agreed that the noise level in their area was occasionally too high and they discussed how, as members of a team, they could create a quieter learning atmosphere. Because these teachers worked together, they were motivated to help each other—to give feedback that would improve not only their individual teaching performance but the overall atmosphere of their classroom.

One teacher pointed out that a major difference between criticism during collegial evaluation and evaluations by an administrator was "the way criticism was phrased." We were continually impressed by the tact and diplomacy exhibited in the conferences. Criticisms were frequently presented as suggestions for alternative techniques. In one teacher's words, "Instead of having someone say, 'you should do this', a colleague was more likely to say, 'something that worked well for me was this technique.'" This approach not only was less threatening but was perceived as more legitimate. If the technique worked for a colleague, it was worth a try.

The interns' conferences emphasized diagnosis rather than specific recommendations. They spent more time and effort analyzing teaching strengths and weaknesses than the elementary school teachers did. Perhaps because of their relative inexperience, they did not have as many concrete suggestions to offer one another and instead devoted some time at each conference to brainstorming alternative teaching strategies.

Collegial evaluation provided positive reinforcement as well as constructive criticism. Suggestions for improvement were balanced with praise for effective teaching. Praise seemed to fill a very great need. As one teacher said, "When your colleague praises you, it means so much." Praise improves teaching by reinforcing successful practices, thus encouraging their frequent use. In school, teachers rarely receive praise from their colleagues because they are not observed or evaluated by them. Though the value of positive reinforcement in motivating pupils is universally recognized, this practice has seldom been extended to teachers--in spite of the fact that the importance of teachers' job satisfaction and faculty morale has long been recognized by teachers and administrators alike.

The feedback given in the conferences encompassed virtually every aspect of classroom activity. Teachers learned not only about their own performance but about the overall climate of their classroom. For example, one intern noted, "There was a warm, cooperative atmosphere in this classroom. It was created by allowing student work groups to sit together on pillows on the floor and emphasizing the importance of group evaluation for the task." Another intern summarized his feeling for a class by telling his partner, "People are noisy; that doesn't bother me. They are talking, getting excited, and having fun." On a more critical note, an art intern told his partner that clean-up period was "utter chaos" and suggested that students be assigned responsibilities for cleaning up after themselves.

Teachers also reported learning more about the behavior of particular students. One observer said of a self-directed project, "The autonomous kids go directly to work, but those who need a lot of teacher direction and support are left out." During a classroom discussion session, another observer noted, "While most students seem to be involved, a few appear to be untouched by the discussion." And during a lecture presentation another observer said, "A couple of students did not understand; they needed extensive clarification." These comments became catalysts for discussion in the conference. The observed teacher wanted to know which students were not autonomous, which were untouched by the discussion, and which needed further clarification. The partners then discussed ways to overcome these problems.

Some of the observations focused on problems of classroom discipline. Classroom control was more frequently discussed in conferences by interns than by teachers. Throughout the evaluation process, interns helped one another identify which students were creating problems and what might be done to improve classroom order. For example, one intern learned that "a small group of boys in the back are goofing off." Following the conference this small group was broken up and dispersed throughout the classroom.

After specific discipline problems had been openly discussed in the conferences, both interns and teachers often took steps to solve them. Overlooking a particularly noisy student is difficult when a colleague has identified the problem through systematic evaluation and provided a justification for action. For example, many interns reported a reluctance to openly chastise their students. They feared that any display of authority would squash independence or creativity, or perhaps more important that it would jeopardize their students' affection for them. But when a colleague says that a certain student is testing the limits of tolerance (and what's more, that the same student creates a similar problem in his or her own classroom), a teacher feels more justified in trying to find sound teaching techniques to bring that student into line.

Understandably, much of the feedback exchanged during conferences focused on the teacher's behavior in the classroom. Some discussions were directed at subject-matter presentation. Teachers gave each other useful information about the quality of materials used in lessons, the appropriateness of the language used in classroom presentations, the clarity of objectives and direction, and specific techniques for making their lessons more interesting. These comments ranged from general observations, such as "The material is going over the kids' heads," to more specific one, such as "Your explanation of chromatic half steps was a little complicated." Similarly, the suggestions for improvement ranged from general ones concerning the teacher's overall performance, such as "You should take at least a half hour to present material you are now covering in ten minutes," to very specific ones, such as "Why not give each student a copy of the keyboard to follow along during your explanation of chromatic half steps?"

The conferences also provided a forum for discussing teacher-student interaction, which was a matter of great concern to the participants, judging by both the criteria they chose for observing and the feedback they gave during conferences. A common observation was that a certain student or group of students was ignored. Many teachers wanted feedback concerning whether they used eye contact with everyone in their room, whether they called on different pupils rather than continually selecting the same ones, and whether they gave equal attention to students. One teacher learned that though she was successful in finding occasions to talk with all of her students individually about their art projects, most of her remarks were negative. In the conference her partner suggested that "students should get more reinforcement on the positive aspects of their work." Teachers continually praised one another for using positive reinforcement.[1] As one said, "You gave lots of 'warm fuzzies' this morning and it meant a lot to the kids."

On a more procedural note, participants found that holding conferences no more than two or three days after observations improved the quality of feedback. Similarly, the observation form (where ratings and comments on the colleague's performance are written) was more useful if it was completed immediately after observing. But most important, teachers reported that the quality of their conferences ultimately depended on the willingness of the partners to be reasonably honest with one another.

---

[1]Teachers rarely told one another to be more critical of their students' work or to develop higher expectations for their students, either individually or as a class. They seemed to believe that each student should receive a lot of teacher warmth and approval regardless of his academic performance. We believe that this approach has serious flaws. Other research shows that students develop greatly inflated opinions of their academic skills in classrooms characterized by strong and uncritical teacher approval. Overstressing warmth and praise may have negative consequences, since it can lead students to have totally unwarranted beliefs about their academic skills. G.C. Massey, M.V. Scott, and S.M. Dornbusch, Racism without Racists: Institutional Racism in Urban Schools, Occasional Paper No. 8 (Stanford, Ca: Stanford Center for Research and Development in Teaching, 1975), pp.7-10. Reprinted from The Black Scholar, 7, No.3 (November 1975), pp. 10-19.

Self-Assessment and Student Questionnaire

Following the structure of our collegial evaluation program, several of those who participated in the pilot test distributed the student questionnaire to their classes and completed the self-assessment form as part of the evaluation process. The teacher questionnaire contains items parallel to the student questionnaire. These allow teachers to identify similarities and differences in their perceptions of themselves and their students' perceptions. For example, the teacher responds to the question, "How often do you encourage students to ask questions when they don't understand what's going on?" Students answer the similar question, "When you don't understand what's going on in this class, how often are you encouraged to ask questions?" Like the teacher, students use a five-point scale which ranges (for this question) from "always" to "never." After combining the student responses and computing a classroom average, the teacher can discover the level of agreement between his self-assessment and his students' assessment. Moreover, by looking at the distribution of responses, a teacher might find that some students "never" feel encouraged to ask questions, even though most students "usually" do. Both the classroom average and the distribution thus provide interesting and useful kinds of information.

The contribution of these questionnaires to the evaluation process was summarized by one teacher:

> I believe that the student questionnaire was extremely valuable in providing information that I myself or a third person could not possibly provide adequately or accurately. The specific kinds of questions deal with those problems that cannot be readily observed. They focus on those students' personal and academic needs that are basic to learning.

One of the most striking results of the pilot test was the high level of agreement between teachers and students as shown by responses on their questionnaires. This similarity was not anticipated by the teachers. One teacher remarked, "I was very surprised to find that my own perceptions agreed fourteen out of twenty-one times (over 66%) with the average of the students. I think this proved that even though my class may not be the greatest one in the world, my students and I certainly agree on what it is." Another teacher said, "The questionnaires indicate that I

have a realistic understanding of my students' feelings toward the class and myself as a teacher."

Despite the general agreement, there were several items on the questionnaire that produced substantial disagreement between teachers and their students. These findings raised new questions and prompted teachers to investigate the underlying reasons for the discrepancy. For example, one teacher was surprised to find that on the average her students felt classwork was "usually" too fast and difficult. Her first interpretation was that she had overestimated her students' abilities. After looking more closely at the distribution of responses, she saw that almost as many students felt the work was "just right" as felt the work was "much too difficult." The second interpretation focused on the diversity of student ability in the classroom. To improve her teaching, she began to individualize instruction so that all of her students would be able to do some things well.

General disagreement was produced between the intern teachers and their high school students by another interesting question: "How important to you is having the teacher like you?" Secondary students rarely reported that this was either "extremely" or "very" important. The secondary interns seemed a little hurt and surprised by their students' indifference. This finding generated a very fruitful discussion among interns. It led to admissions that they were probably upset by this student report because they wanted so much to be liked by their own students. They had just assumed that liking was reciprocal. They confided to one another that wanting to be liked sometimes interfered with their better judgment as teachers. This conclusion was incorporated into their overall plans for improvement.

By comparison, elementary teachers were a little overwhelmed at their students' rating of their teacher's importance in their lives. Almost all elementary students said it was "extremely important" to be liked by their teacher. Of course, these veteran teachers had suspected that their students wanted their affection, but they had not known how strong or how widespread this feeling was. Such unanimity in their students' responses made them sensitive to a number of related behaviors in the classroom.

For example, after reviewing the questionnaire but prior to observation, one teacher noted about another, "Those kids are always touching you, and you never fail to respond."

In addition to insights gained from students' responses on each item of the questionnaire, teachers discovered that examining the responses on several items at once sometimes revealed interesting patterns. For example, one teacher discovered that her students reported being more confused than she had suspected. They agreed that the teacher's directions were unclear and that they were seldom encouraged to ask questions. She felt that their confusion might be alleviated if she took measures to clarify her directions and encouraged them to ask questions whenever they were confused.

Although anonymity was ensured on the student questionnaire, teachers and interns spent a lot of time guessing which students had given certain responses. The elementary teachers, who knew their students much better than the interns, seemed confident of their ability to make these guesses. When one student responded that he "never received good grades" even when he did "good work," the teacher said, "I know who that is, and he's right. We've got to start giving him some rewards for his efforts." The teacher was confident that this was the same student who responded that the teacher never let him know when he was doing "good work."

The participants agreed that maintaining anonymity was important if they wanted honest responses from students, but one lamented that "it would be valuable to know a particular student whose answers were radically different. It may be that this student is having difficult problems that I have overlooked or that are not obvious to me, and I would want to give him the special help that might be needed."

In the pilot test, one of the interns did a fine job of developing his own student questionnaire. He wanted to obtain specific information about his skills as a choir director. He learned that his conducting was "fairly easy to follow," but almost half of his students felt that he "stayed on one piece of music too long." Most of the choir liked the music "O.K.," with just a few liking it "a lot" or "not much." Only two students thought he looked like a "madman" when conducting. These items

provided an excellent supplement to the more general student questionnaire.

Student questionnaires provide teachers with information they cannot obtain elsewhere. Only students can tell a teacher whether or not they are interested and comfortable in the classroom. The problems students perceived were translated into specific criteria for the teacher's colleague to observe and were discussed in the conferences. The student assessment was a very valuable input that the teachers took into account in assessing their strengths and weaknesses and making plans for improvement.

Self-Assessment on Selected Criteria

In addition to the teacher questionnaire, participants completed a self-assessment form based on the criteria they had selected jointly with their partners. After their teaching was observed, this self-assessment could be compared with the observation form to help focus the conference on areas for improvement. Overall, participants were usually much more critical of themselves, both in ratings and in negative comments, than their colleagues were. They generally agreed with their partners' observations on areas of weakness, and most spent their conference in swapping ideas for improvement rather than in resolving disagreements.

A colleague's agreement was helpful in legitimatizing a teacher's perception of her strengths and weaknesses. For example, one teacher commented, "In discussion, I tend to rely on the same students who always have the answers, and I do not phrase open-ended questions to include everyone." When her colleague noted that "two boys spoke often, a few girls spoke occasionally, but no one else entered the discussion," her self-assessment was confirmed. A good part of their first conference focused on how she might increase student participation. In the second observation her colleague noted that "the discussion included more students and some who had not previously participated. You praised the newcomers-Good."

In her self-assessment another teacher noted a need for "some improvement" in lectures because she "relied too heavily on note cards." During the first observation her colleague identified the same area: "The organization and sequence of the lesson is good, but you occasionally stopped to

refer to notes." At the second observation the problem was not as severe and the colleague observed, "You relied on notes much less."

Of course, not all of the problems were so easily remedied. In a self-assessment one teacher reported the need "to project my voice." Her colleague noted, "Teacher's quiet voice tends to trail off" on the first observation form. In the second observation period the colleague reported, "Teacher's voice does not carry above sound of the slide projector." This is clearly a problem that needs to be addressed in that teacher's improvement plan.

The Improvement Plan

Developing a plan for improvement is the most important step of the collegial evaluation process. But the quality of each teacher's plan depends on how well the other steps have been carried out. The plan for improvement is formulated in a final "wrap-up" conference between the two partners. Each teacher integrates all the information he or she has received from self-assessment, student questionnaires, and peer evaluation, and presents his partner with a composite list of strengths and weaknesses. Together the teachers decide on the specific strategies each will use to improve their teaching performance in areas of weakness. In addition, they determine how they will evaluate the results of these strategies. Finally, they identify any resources they will need to carry out their improvement plan.

In our pilot test of collegial evaluation, the improvement plans spanned the whole range of teaching activities: presentation of subject matter, classroom control, motivation, student interest and involvement, positive reinforcement, and classroom organization and atmosphere. The improvement plans were based on evaluations that showed a remarkable amount of agreement between the teachers themselves, their colleagues, and their students. In most cases a teaching weakness identified by one of these sources was corroborated by the others.

For example, one teacher listed as an evaluation criterion, "Do not ignore any segment of the class concerning questions or needs--give attention equally." On the student questionnaire several students reported that

they were "seldom" or "never" encouraged to ask questions in class. On the basis of classroom observation the teacher's partner noted: "The less capable students are not involved, especially those at the back." As part of the plan for improvement, the teacher specified, "With the help of my peer, I will first identify those students whom I have ignored. I will make a point of talking to each of them every day. I'll keep a check list to make sure I spend some time with each of these children." Another teacher developed a plan to deal with a similar interaction problem in a different way. To encourage the nonparticipators at the back, she decided to rearrange the class and move the pupils at the back into the first two rows. She also said that she would "give those individuals who have not been participating responsibility for explaining things to the class and helping others with their work."

An intern chose as an evaluation criterion, "I present subject matter at a level appropriate to student ability." He was perplexed when most of his students reported on the questionnaire that they were confused by his explanations. Then his peer commented, "You use a lot of terms which go way over some of these kids' heads." In his improvement plan this intern listed a number of specific strategies to overcome the problem. Among these were: "I will try to define clearly all new terms which I use in class and be more careful to write these terms and their definitions on the board. I'll use pretests to determine pupil knowledge in the subject area. For those who do well on these tests I will design self-directed projects. This will leave me free to spend more time with the slow-achievers."

Some of the improvement plans called for relatively minor changes; others envisioned a major reorganization of the classroom and substantial changes in teacher behavior. Two of the elementary school teachers felt that they both needed to maintain a quieter learning environment. Such a concern is not atypical in open-space classrooms. After observing one another, they discovered that the noisiest time of the day came when they grouped their students by ability in math and language arts. The noise came from the "low ability" youngsters, and it prevented them and others from concentrating. As part of their improvement plan, the teachers

decided that the next year they would experiment with more heterogeneous groups.

Many of the identified weaknesses were not so difficult to remedy. For example, one art teacher, concerned about giving appropriate positive reinforcement for good work, benefited from his colleague's observation that he did not have any student work displayed in the classroom. He planned to "reserve a large space in the art room, school library, and hall display cases for the exhibition of student work." Another intern, whose problem was that he never had time to finish his lesson, decided to save a few minutes each period by letting students distribute and collect classroom materials rather than doing it himself.

For each of the specific strategies, teachers were asked to determine how they would assess their progress. Plans for assessment were as varied as improvement strategies. Teachers planning to improve their presentation of subject matter often relied on student cognitive outcomes as a measure of their success. The teacher mentioned above, who planned to explain and define new terms more carefully, listed as one indicator of progress the number of times students used the new terms in their essays.

Several teachers decided to use the student questionnaire as a post-test device to assess their improvement. Comparing the student response before and after the improvement plan was put into effect would help them assess their progress in such areas as motivating students, evaluating them, presenting material clearly, individualizing subject matter, displaying interest in students, and developing material appropriate to the students' level.

Almost all of the teachers planned to use collegial observation and conferences as a method of assessing their improvement. Many had already set up times to begin another round of observations with their colleagues. Others decided to change partners. The specific strategies for improvement would suggest new criteria for the next round of observations. One of the most gratifying results of the pilot test was that many of the participants considered our collegial evaluation program so useful that they planned to extend it throughout the school year. As one teacher said,

> I need to have this kind of collegial evaluation on a regular basis. If my colleague evaluated me

> throughout the year, she would have an understanding of the trends in my teaching and in a particular class and the evaluation would be even more helpful. She would be able to detect subtle problem areas that I may not be aware of. I could do the same for her and also continue to learn a lot by observing another teacher at work.

Conclusions

We began this discussion by criticizing traditional approaches to teacher evaluation and advocating collegial evaluation as an alternative. We summarized research revealing that teacher evaluation programs are all weak in one or more steps of the evaluation process. According to teachers and administrators we have interviewed, criteria for observation are usually vague or unknown, observations are infrequent, useful feedback is rare, and plans for teacher improvement are almost nonexistent. The experiences of teachers in the pilot test of our collegial evaluation program gave us some evidence for assessing this approach and comparing it with more traditional methods of evaluating teachers.

Most important, we learned that teachers can and will help each other perform better on their jobs. We also learned that teachers will take students' assessments of their teaching seriously and use them in developing plans for improvement.

We found that the most difficult step of our program was selecting criteria to serve as a basis for evaluation. But most teachers did select some criteria that were specific, observable, and meaningful to them. We also learned that thinking about their criteria helped teachers assess not only where they might need to improve but what their goals as teachers were.

We emphasized that the steps of the evaluation program are interdependent and that a weakness in any one of them would diminish the program's usefulness. This was especially apparent in reviewing improvement plans. If the criteria were specific, observable, and meaningful, if the observer was attentive and carefully reported observations to his or her colleague, and if the feedback exchanged was complete and honest, then the improvement plan generated by the pair of teachers was a thoughtful and practical blueprint for professional growth. The message is clear; teachers cannot

participate in this program in a half-hearted manner. If they are to use it as a means for improving their teaching, they must commit themselves to doing a thorough and careful job at every step.

Does collegial evaluation work? We believe the answer is yes. Based on our pilot test we have concluded that collegial evaluation is a useful approach to teacher evaluation in schools. On the whole, teachers reacted favorably to collegial evaluation, adapted the program to fit their unique circumstances, and gained new ideas for improving their teaching.

References

Dornbusch, S. M., and Scott, W. R. Evaluation and the Exercise of Authority. San Francisco: Jossey-Bass, 1975.

Marram, G. D., Dornbusch, S. M., and Scott, W. R. The Impact of Teaming and the Visibility of Teaching on the Professionalism of Elementary School Teachers. (Stanford Center for Research and Development in Teaching, Technical Report No. 33) Stanford University, December 1972.

Thompson, J. E., Dornbusch, S. M., and Scott, W. R. Failures of Communication in the Evaluation of Teachers by Principals. (Stanford Center for Research and Development in Teaching, Technical Report No. 43) Stanford University, April 1975.